

Research Article

Open Access

Predicting Under Nutrition Status of Under-Five Children Using Data Mining Techniques: The Case of 2011 Ethiopian Demographic and Health Survey

Zenebe Markos¹, Feleke Doyore^{2*}, Martha Yifiru³ and Jemal Haidar³

¹Dilla University, Information and Communication Directorate, Ethiopia

²Department of Public Health, Wachemo University, Ethiopia

³Department of Health Informatics and Data Mining, Addis Ababa University, Ethiopia

Abstract

Background: Under nutrition is one of the leading causes of morbidity and mortality in children under the age of five in most developing countries including Ethiopia. The main objective of this study was to design a model that predicts the nutritional status of under-five children using data mining techniques.

Methods: This study followed hybrid methodology of Knowledge Discovery Process to achieve the goal of building predictive model using data mining techniques and used secondary data from 2011 Ethiopia Demographic and Health Survey (EDHS) dataset. Hybrid process model was selected since it combines best features of Cross-Industry Standard Process for Data Mining and Knowledge Discovery in Database methodology to identify and describe several explicit feedback loops which are helpful in attaining the research objectives. WEKA 3.6.8 data mining tools and techniques such as J48 decision tree, Naïve Bayes and PART rule induction classifiers were utilized as means to address the research problem.

Result: In this particular study, the predictive model developed using PART pruned rule induction found to be best performing having 92.6% of accurate results and 97.8% WROC area. Promising result has been achieved from the rules regarding nutritional status prediction.

Conclusion: The results from this study were encouraging and confirmed that applying data mining techniques could indeed support a predictive model building task that predicts nutritional status of under-five children in Ethiopia. In the future, integrating large demographic and health survey dataset and clinical dataset, employing other classification algorithms, tools and techniques could yield better results.

Keywords: Predictive modeling; Nutritional status; Children; Data mining; EDHS dataset

Introduction

Good nutrition is an essential component of good health. Nutrition is at the heart of most global health problems – especially in the area of child survival where child under nutrition is an underlying cause of more than one-third (3.5 million) prevalence of all child deaths under the age of five in developing countries. Of the 112 million underweight children and 178 million children who suffer from stunting, 160 million (90%) live in just 36 developing countries, constituting almost half (46%) of the cases [1].

Malnutrition is a known contributing factor to disease and death. Worldwide, over 10 million children under the age of five years die every year from preventable and treatable illnesses despite effective health interventions [2]. One in four of the world's children are stunted [3]. At least half of these deaths are caused by malnutrition. In addition, malnourished children that survive are likely to suffer from frequent illness, which adversely affects their nutritional status and locks them into a vicious cycle of recurring sickness, faltering growth and diminished learning ability [4].

Nutritional status is an outcome and impact indicator when assessing progress towards achieving the Millennium Development Goals (MDGs). Marked differences, especially with regard to height-for-age and weight-for-age, are often seen among different subgroups within a country. Child nutritional status is related with MDGs special Goal 4 of 2/3 child mortality reduction in 2015. Improving child nutrition is a key to achieving this goal. Child mortality is

deeply interlocked with all the other MDGs [1]. Each one is a major contributor to poor and dangerous living conditions for children.

Federal Ministry of Health (FMOH) and Central Statistics Agency of Ethiopia (CSAE) in collaboration with non-governmental organizations collected large volume of dataset to identify children nutritional factors and reported nutritional status of children using SPSS for analysis with limited tools [5]. These limited attributes can be improved using a technology that has the capacity to extract hidden useful information from such type of data through data mining. Data mining is a crucial step in discovery of knowledge from large datasets. In recent years, data mining has found its significant hold in every field including health care. Mining process is more than the data analysis which includes classification, clustering, and association rule discovery. It also spans other disciplines like data warehousing, statistics, machine learning and artificial intelligence.

***Corresponding author:** Feleke Doyore, Department of Public Health, Lecturer and Faculty Dean, Faculty of Medicine and Health Science, Wachemo University, Hossana, Ethiopia, Tel: +251916291489; Fax: +251932685424; E-mail: feledoag@yahoo.com and fdoyore@yahoo.com

Received February 11, 2014; **Accepted** February 17, 2014; **Published** February 24, 2014

Citation: Markos Z, Doyore F, Yifiru M, Haidar J (2014) Predicting Under Nutrition Status of Under-Five Children Using Data Mining Techniques: The Case of 2011 Ethiopian Demographic and Health Survey. J Health Med Informat 5: 152. doi:10.4172/2157-7420.1000152

Copyright: © 2014 Markos Z, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

It is therefore reasonable to study that what attribute values are affecting nutritional status and develop a model that assists in predicting future interventions based on the values of significant attributes identified. The ability of the tools and algorithms of data mining to deal with datasets characterized by thousands of instances and high dimensionality (large number of attributes) coupled with the understandability of models produced at the end and their ease of use would make data mining suitable for this study. Now a day, data mining technology is being used as a tool that provides the techniques to transform these mounds of data into useful information which in turn enables to derive knowledge for decision making. In addition to this, it is better to study using large datasets and attributes with advanced technology called data mining.

The main purpose of this study was to apply data mining techniques for extracting hidden patterns which are significant to predict nutritional status of under-five children from 2011 EDHS dataset. To achieve this goal, the following research questions were formulated for investigation.

- What are the optimal determinant factors that lead to child under nutrition in Ethiopia?
- Which predictive modeling algorithms are suitable for determining nutritional status of under-five children?

Furthermore, the findings of this study will enable data miners, message developers, medical professionals, researchers and policy makers used as baseline data to design appropriate and effective intervention.

Methods (Research Design/Data Mining Modelling)

The study followed hybrid methodology of Knowledge Discovery Process (KDP) to achieve the goal of building predictive model using data mining techniques. Hybrid process model was selected since it combines best features of CRISP-DM and KDD methodology to identify and describe several explicit feedback loops which are helpful in attaining the research objectives. Hybrid methodology basically involves six steps (Figure 1).

The Weka GUI chooser

Provides a starting point for launching Weka’s main GUI applications and supporting tools. It includes access to the four Weka’s main applications: Explorer, Experimenter, Knowledge Flow and Simple CLI.

Classifier accuracy (performance evaluation) measures

In order to minimize the bias associated with the random sampling of the training and test data samples k-Fold Cross Validation was adopted. In k-fold cross-validation, the initial data are randomly partitioned into k mutually exclusive subsets or “folds,” D1, D2, : : : , Dk, of approximately equal size.

10-Fold Cross Validation: In 10-fold cross validation, the complete dataset is randomly split into 10 mutually exclusive subsets of approximately equal size. Each time it is trained on nine folds and tested on the remaining single fold. 10-fold cross validation does not require more data compared to the traditional single split (2/3 training, 1/3 testing) experimentation.

Confusion matrix

Useful tool for analyzing how well classifier recognized the classes. An entry, CM_{ij} in the first m rows and m columns indicate the number of tuples of class that were labeled by the classifier as class j [6].

Receiver operating characteristic curve

To test which classifier is highly significant for a given subject is determined by ROC analysis and it becoming widely used tool in medical tests evaluation [7].

Understanding and Pre-processing 2011 EDHS Dataset

The raw data description

The source data employed for this research purpose is 2011 EDHS dataset. This dataset is collected from 2006/2007-2010/2011. The 2011 EDHS was conducted under the support of the MOH and CSA. The census is conducted in every five years intervals. The primary objective

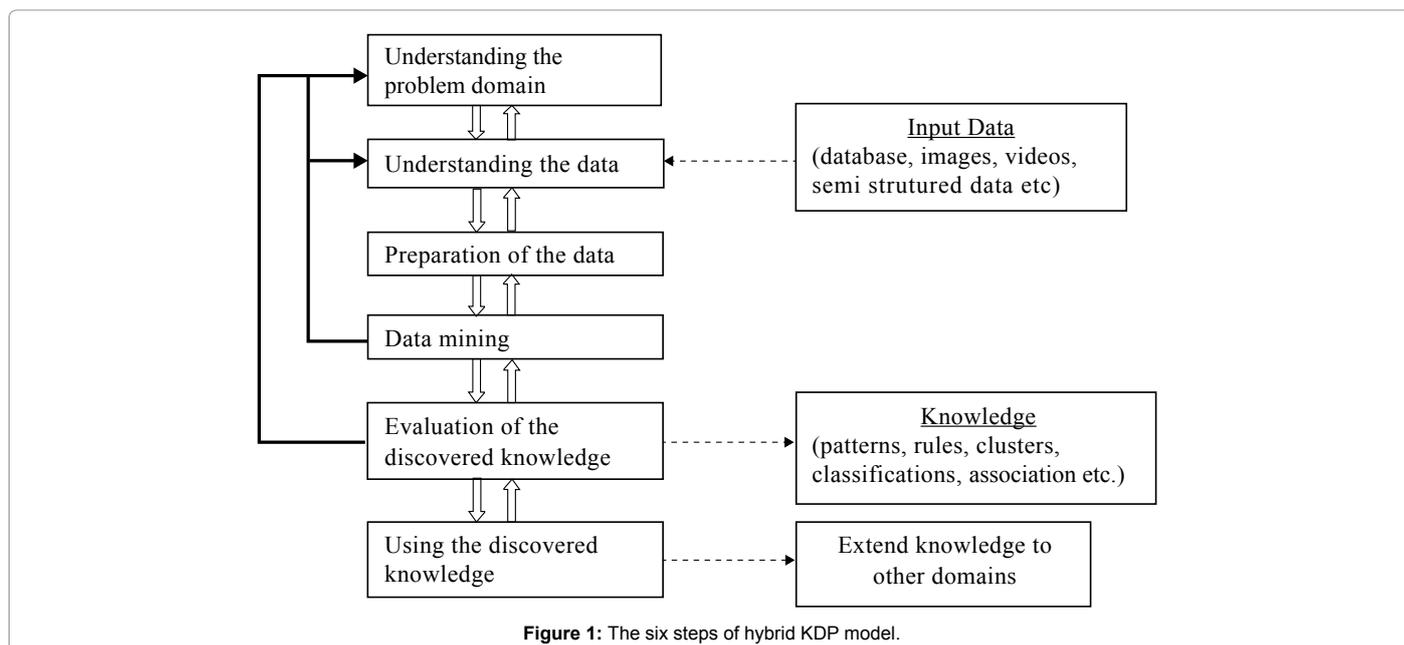


Figure 1: The six steps of hybrid KDP model.

of the 2011 EDHS was to provide up-to-date information for policy makers, planners, researchers and programme managers, which would allow guidance in the planning, implementing, monitoring and evaluating of population and health programmes in the country. The various sector development policies and programmes assist and monitor the progress towards meeting the Millennium Development Goals [5].

The 2011 EDHS collected information on the population and health situation which covers family planning, fertility levels and determinants, fertility preferences, infant, child, adult and maternal mortality, maternal and child health, nutrition, malaria, and women’s empowerment.

The 2011 EDHS dataset has eight data (i.e, children’s records, birth records, couple’s records, HIV test record, household member records, individual records and male records) in SPSS file format. Each data include other attributes such children’s data. In children’s data, there are household attributes, male attributes, birth attributes and couples attributes with the total of 920 attributes. The 2011 EDHS is a nationally representative survey of 11, 654 instances on children collected in order to classify nutritional status, anemia level, and others in the study from women age of 15-49, men age of 15-59 years and under five year children on 920 attributes. This sample provides estimates of health and demographic indicators at the national and regional levels, and for rural and urban areas. Among all under-five survey data, child data attributes such as sex of child, age of child, height of child, weight of child, Height for Age Z-score (HAZ), Weight for Age Z-score (WAZ), Weight for Height Z-score (WHZ), anemia level (hemoglobin level), etc are included. Mother’s background characteristics are mother’s age, region, place of residence, literacy, BMI, religion, wealth index, mother’s education, occupation, hemoglobin level (anemia), etc (Table 1). On the other hand, the data also included household (HH) characteristics such as, age of -HH, sex of HH, education of HH, occupation of HH, etc.

From original children’s data other’s attributes and unrelated attributes with nutritional status of under-five children removed. Finally the selection of the dataset is performed by the help of literature and domain experts. List of source dataset, 2011 EDHS dataset, before preprocessing is depicted in result part.

Data understanding: This phase mainly focuses on creating a target dataset with selected sets of variables that is relevant to the discovery process. Without understanding the existing data, it is difficult to draw the target dataset from the original since the world data is unclean and not appropriate at the source to run mining process [8].

The original dataset from SPSS is exported to excel file because Weka data mining tool does not accept SPSS format and whose size amounted to 14.6 MB before any processing activity was done on it.

MOTHER’S AGE			
N	Category	Frequency	Percent (%)
Valid	15-19	393	4.1
	20-24	1893	19.7
	25-29	3019	31.4
	30-34	1996	20.8
	35-39	1487	15.5
	40-44	615	6.4
	45-49	204	2.1
Missing		0	0.00
Total		9607	100

Table 1: Summary of Mother’s Age Attribute.

Under-five data which found in electronic format has 11,654 instances and 920 attributes. This 920 attributes are not only on nutritional status but on HIV, vaccination, breast feeding, child preference, nutrition (under-five, adult and women), etc.

The entire attributes in the original dataset were not concerned for this experimentation. Thus, only relevant attributes were considered so as to achieve the objective of the study. From total of 920 attributes which were found in under-five children records, 44 attributes which related with nutritional status were selected. From total 44 attributes, 14 repeated attributes, 8 least important attributes and 6 attributes those have more than 50% missing values were minimized. Finally, source data was not labeled or clustered by nutritional status (new class). Based on the 2006 WHO children multicenter standards, researcher labeled nutritional status [8,9].

Data preparation: These are the most important phases of the data analysis activity which involves the construction of the final data set (data that will be fed into the modeling tool) from the initial raw data. Data preparation generates a dataset smaller than the original one, which can significantly improve efficiency of data mining. This task includes: attribute selection, filling the missed values, correcting errors, or removing outliers (unusual or exceptional values), resolve data conflicts using domain knowledge or expert decision.

Attribute selection: Deciding on the data that was used for the analysis was based on several criteria, including its relevance to the data mining goals as well as quality and technical constraints such as limits on data volume or data types [9]. Therefore, in this study the attributes are selected with the help of domain expert and extensive literature review because taking all the variables in the database we have, feed them to the data mining tool and find those which are the best predictors may be does not work very well. One reason is that the time required to build a model increases with the number of variables. Another reason is that blindly including extraneous columns can lead to incorrect models [10]. Thus, it is necessary to leave out those attributes that are not important for analysis with the help of domain experts and literature review in order to simplify the task of modeling.

The national survey data set obtained contains many attributes. To decide on the relevant attributes for this study, the researcher has discussed with domain experts in the area. The attributes are selected from the five years survey: Mother’s age, Mother’s educational level, Mother’s BMI, Mother’s occupation, Residence, Region, Wealth quintile, Size of child at birth, child’s age, child’s sex, child’s HAZ, child’s WAZ, child’s WHZ, anemia level, total number of children and ever had vaccinated. The final selected attributes were prepared and preprocessed as stated in the following section, before developing the models and explained briefly as follows.

Selection of instances: In addition to the removal of irrelevant attributes which were done based on the attributes; relevance to the prediction of nutritional status of under-five children, instances that deal with nutritional status were selected from the dataset. Out of the 11,654 under-five data, 9,607 remain. As this study uses classification algorithms for the purpose of predictive model building, 846 records without class information are removed from subsequent analyses because of died children. Instance with missing values for outcome class are not useful for predictive model building in data mining because classification algorithms of data mining learn how instances are classified under the different classes. The classes do not exist, means the algorithm learns nothing from these instance. Records without class labels (missing or not entered) were ignored, provided that the data mining task involves classification [11]. Even, from this number, 1201

records are also removed due to over-nutrition. The remaining dataset has 9607 records are distributed in one of the outcome categories (normal, wasted, underweight, and stunted).

Exploratory data analysis: In this section, efforts have been made to present the description of the selected attribute together with the exploratory data analysis performed with the use of frequency tables. With the use of frequency tables, the exploratory data analysis was performed to detect attributes with the missing values and wrong entries.

Mother's age: The age of mothers is classified by five year age groups. This attribute is categorized into seven groups: 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49.

Region and place of residence (urban/rural): The region attribute indicates the location of mothers and child. This attribute contains a total of 11 administrative region of the country (Table 2).

Mother's education: This attribute reveals the level of education of a mother. Mother's education is indirectly related to a child's health. Mother's education is nominal attribute that contains four distinct values (No education, Primary, Secondary, and Higher) (Table 3).

Wealth index: The distinct values of wealth index attributes are poorest, poorer, middle, richer and richest (Table 4).

Total number of children: This is an indicator for number of children born in the family. The distinct values of ever born attributes are 1-2, 3-4, 5-6, and >6 children (Table 5).

REGION			
N	Category	Frequency	Percent (%)
Valid	Tigray	1077	11.2
	Afar	891	9.3
	Amhara	1068	11.1
	Oromiya	1515	15.8
	Somali	755	7.9
	Benishangul-Gumuz	834	8.7
	SNNP	1367	14.2
	Gambela	682	7.1
	Harari	511	5.3
	Addis Ababa	323	3.4
Dire Dawa	584	6.1	
Missing		0	0.00
Total		9607	100

RESIDENCE			
N	Category	Frequency	Percent (%)
Valid	Urban	1544	16.1
	Rural	8063	83.9
Missing		0	0.00
Total		9607	100

Table 2: Summary of Region Attribute.

MOTHER'S EDUCATION			
N	Category	Frequency	Percent (%)
Valid	No education	6695	69.7
	Primary	2446	25.5
	Secondary	307	3.2
	Higher	159	1.7
Missing		0	0.00
Total		9607	100

Table 3: Statistical summary of levels of Mother's Education Attribute.

WEALTH INDEX			
N	Category	Frequency	Percent (%)
Valid	Poorest	2959	30.8
	Poorer	1776	18.5
	Middle	1592	16.6
	Richer	1583	16.5
	Richest	1697	17.7
Missing		0	0.00
Total		9607	100

Table 4: Statistical summary of mother's wealth index attribute.

TOTAL NUMBER OF CHILD EVER BORN			
N	Category	Frequency	Percent (%)
Valid	1-2	2855	29.7
	3-4	2804	29.2
	5-6	1978	20.6
	>6	1970	20.5
Missing		0	0.0
Total		9607	100

Table 5: Statistical summary of total number of ever born children attribute.

MOTHER'S BMI			
N	Category	Frequency	Percent (%)
Valid	<18.5	2520	26.2
	18.5-24.9	6458	67.2
	>=25	604	6.3
Missing		25	0.3
Total		9607	100

Table 6: Statistical summary of mother's BMI attribute.

MOTHER'S OCCUPATION			
N	Category	Frequency	Percent (%)
Valid	Not working	5168	53.8
	Working	4439	46.2
Missing		0	0.0
Total		9607	100

Table 7: Statistical summary of mother's occupation.

SIZE OF CHILD AT BIRTH			
N	Category	Frequency	Percent (%)
Valid	Small	2771	28.8
	Normal	3792	39.5
	Large	3018	31.4
Missing		26	0.3
Total		9607	100

Table 8: Statistical summary of the size of a child at birth.

Mother's BMI: Mother's BMI is an indicator for nutritional status of a mother. The distinct value of mother BMI attribute is under or thin (<18.5 kg), normal (18.5-4 kg) and/or over (>4 kg) (Table 6).

Mother's occupation status: Mother's occupation is an indicator of economic level in relation to nutritional status of a child with distinct values of attribute is not working and working (Table 7).

Size of a child at birth: Size of a child at birth reveals nutritional status of a child in the house hold. The distinct values of this attributes are large (>4 kg), normal (2.5-4 kg) and small (<2.5 kg) (Table 8).

Ever had vaccination: is an attribute used to check that a child is vaccinated or not. The distinct values of this attribute is not vaccinated and vaccinated (Table 9).

CHILD EVER HAD VACCINATION			
N	Category	Frequency	Percent (%)
Valid	No	1975	20.6
	Yes	5236	54.5
Missing		2396	24.9
Total		9607	100

Table 9: Statistical summary of ever had vaccination.

CHILD ANEMIA LEVEL			
N	Category	Frequency	Percent (%)
Valid	Severe	298	3.2
	Moderate	2051	21.3
	Mild	1785	18.6
	Not anemic	4164	43.3
Missing		1309	13.6
Total		9607	100

Table 10: Statistical summary of child anemia level.

SEX OF CHILD			
N	Category	Frequency	Percent (%)
Valid	Male	4888	50.9
	Female	4719	49.1
Missing		0	0.0
Total		9607	100

Table 11: Statistical summary of sex of a child.

CHILD AGE			
N	Category	Frequency	Percent (%)
Valid	<6	1001	10.4
	6-11	967	10.1
	12-23	1775	18.5
	24-35	1876	19.5
	36-47	2054	21.4
	48-59	1934	20.1
Missing		0	0.00
Total		9607	100.0

Table 12: Statistical summary of children's age category.

Child anemia level: This attribute is an indicator for anemia level of a child and an indicator for nutritional status of a child. The distinct values of this attributes are severe, moderate, mild and not anemic (Table 10).

Sex of a child: Sex of a child is nominal attribute with possible values of male and female (Table 11).

Child age: This is the age of a child from 0-59 months. The possible values of this attribute are: <6, 6-11, 12-23, 24-35, 36-47 and 48-59 months (Table 12).

Nutritional status: nutritional status of children under age five is an important outcome measure of children's health. The anthropometric data on height and weight permit the measurement and evaluation of the nutritional status of the children. Initially there was no variable labeled with nutritional status (class) but it contains anthropometric measurements, i.e, Height for Age Z-score (HAZ), Weight for Age Z-score (WAZ) and Weight for Height Z-score (WHZ). Nutritional status classification has been done by the researcher based on WHO Growth Multicenter standards of 2006. Clustered nutritional status attribute values are stunted, underweight, wasted and normal (Table 13).

Data transformation

Once the data has been assembled and major data problems are fixed, the data must still be transformed for analysis [12]. Discretization is the process of converting continuous valued variables to discrete values where limited numbers of labels are used to represent the original variables. The discrete values can have a limited number of intervals in a continuous spectrum, whereas continuous values can be infinitely many [13]. Here discretization is made on three anthropometric indices (HAZ, WAZ and WHZ) based on 2006 WHO Growth Multicenter standards and size of child at birth

Discrediting the values of HAZ attribute: The attribute HAZ index provides an indicator of linear growth retardation and cumulative growth deficits in children. Final HAZ attribute discredited categories are <-2SD to -3SD moderately malnourished, and <-3 severely malnourished. Finally these two values discredited into one value, stunted. -2SD to 2SD (normal) and >2SD (over) (Table 14).

Discrediting the values of WAZ attribute: Attribute WAZ is a composite index of height-for-age and weight-for-height. It contains chronic and acute malnutrition. Final WAZ attribute discredited categories are <-2SD (underweight), -2SD to 2SD (normal) and >2SD (over) (Table 15).

NUTRITIONAL STATUS OF UNDER-FIVE			
N	Category	Frequency	Percent (%)
Valid	Normal	4713	49.1
	Stunted	2339	24.3
	Underweight	1449	15.1
	Wasted	1106	11.5
Missing		0	0.0
Total		9607	100

Table 13: Statistical summary of nutritional status attribute.

HAZ			
N	Category	Frequency	Percent (%)
Valid	<-2SD	4070	42.37
	-2SD-2SD	5262	54.77
	>2SD	275	2.86
Missing		0	0.0
Total		9607	100

Table 14: Statistical summary of HAZ attribute.

WAZ			
N	Category	Frequency	Percent (%)
Valid	<-2SD	2884	30.0
	-2SD-2SD	6663	69.4
	>2SD	60	0.6
Missing		0	0.0
Total		9607	100

Table 15: Statistical summary of WAZ attribute.

WHZ			
N	Category	Frequency	Percent (%)
Valid	<-2SD	1128	11.7
	-2SD-2SD	8315	86.6
	>2SD	164	1.7
Missing		0	0
Total		9607	100

Table 16: Statistical summary of WHZ attribute.

Discrediting the values of WHZ attribute: The WHZ index measures body mass in relation to body height or length; it describes current nutritional status. Final WHZ attribute discretized categories are <-2SD (Wasted), -2SD to 2SD (Normal) and >2SD (Over) (Table 16).

Discrediting the values of size of child at birth attribute

For easy interpretation, high level concept is generated for size of child at birth attribute. The values are very large and larger than average is more than 4 kg, average (normal) is equal to 2.5-4 kg and smaller than average and very small is less than 2.5 kg. This is done by defining a portion of the values through explicit data grouping as presented in (Table 17). It shows that discretization of size of a child at birth. Except second one, Normal, the first two values merged into large and last two values merged into small. At the beginning of this chapter the dataset acquired as an information source is described. Since then, different activities were performed on the dataset with the objective of making it suitable for the data mining algorithms and producing representative model. Very large numbers of instances were removed and large numbers of attributes are removed. The final summary of the dataset ready for experiments is shown in (Table 18).

Experimentation and evaluation of discovered knowledge

Experimental design: Cleaned dataset is used for predictive model building. All the experiments that are discussed in the subsequent sections are carried out using 9607 instances and 17 attributes. The attribute set includes HAZ, WAZ, WHZ, MOTH AGE, REGION, RESIDENCE, MOTH EDUC, WEALTH INDEX, MOTH BMI, MOTH OCCUP, TOTAL, SEX, CHILD AGE, SIZE, CHILD ANEM, EVER HAD VAC and NUT STATUS. The last attribute in the list represents the class attribute which is mandatory in developing predictive models i.e. the dependent variable in statistics language.

In order to build predictive models for nutritional status, three different data mining algorithms were applied. More specifically, J48 decision tree, Naïve Bayes and PART rule induction are the algorithms with which predictive model building experiments are conducted. In 10 fold cross validation, one option in Weka for the purpose mentioned; the dataset is split into 10 equal parts. The algorithms used for predictive model building experimentations are found in Weka 3.6.8. The prepared dataset is saved using CSV file format. Then, this file is imported to Weka.

A dataset is imbalanced if the classification categories are not approximately equally represented. Often real-world data sets are predominately composed of “normal” examples with only a small

Old Values	New Values
Very Large	Large
Larger than Average	
Average	Normal
Smaller than Average	Small
Very Small	

Table 17: Size of child at birth attributes values generated by explicit data grouping

Parameters	Original dataset	Target dataset
Total Number of Records	11,654	9,607
Total Number of attributes	920	17
File Format	SPSS 16.0	.xls
Size of Data	14.6 MB	2.42 MB
		1.14 MB

Table 18: Summary of the selected dataset.

Class	Before SMOTE	100% SMOTE	200% SMOTE	300% SMOTE	400% SMOTE
Normal	4713	4713	4713	4713	9426
Wasted	1106	2212	2212	2212	2212
Underweight	1449	1449	2898	2898	2898
Stunted	2339	2339	2339	4678	4678

Table 19: Imbalanced classes before and after SMOTE.

Rank#	Name Of Attributes	Information gain	Rank#	Name Of Attributes	Chi Square Attribute Eval
1	HAZ	0.73816	1	WHZ	14268.19418
2	WHZ	0.60394	2	HAZ	11368.24155
3	WAZ	0.56478	3	WAZ	8751.37242
4	CHILDAGE	0.08292	4	CHILDAGE	1577.62869
5	REGION	0.047	5	REGION	938.16191
6	WEALTHINDEX	0.03077	6	WEALTHINDEX	632.96984
7	MBMI	0.02089	7	MBMI	490.24783
8	MATHEDUC	0.02094	8	MATHEDUC	415.93063
9	RESIDENCE	0.01894	9	RESIDENCE	394.25737
10	SIZE	0.01603	10	SIZE	321.46004
11	CHILDANEM	0.01447	11	CHILDANEM	293.00299
12	EVERHADVAC	0.00805	12	EVERHADVAC	163.18881
13	MOTHAGE	0.00728	13	MOTHAGE	145.82242
14	MOTHOCUP	0.00708	14	MOTHOCUP	141.46341
15	TOTAL	0.00308	15	TOTAL	62.28584
16	SEX	0.00139	16	SEX	27.85793

Table 20: Attributes evaluation by Information Gain Attribute Eval and Chi Square Attribute Eval.

percentage of “abnormal” or “interesting” examples. There is a special technique called SMOTE to solve such type of problems in dataset (Table 19). Over-sampling the minority class and under-sampling the majority class can achieve better classifier performance (in ROC area) than varying the loss ratios.

Selecting and evaluating the attributes

To do this, “Info Gain Attribute Eval” and “Chi Square Attribute Eval” methods were used to assign a worth to each subset of attribute by searching ranker style in the Weka (Table 20).

Algorithm classifier parameters

J48 Classifier parameter options: During experimentations, all the default parameters that already set in the Weka were used except that the unpruned ‘False’ was changed to unpruned ‘True’ during J48 unpruned tree generation in order to experiment the model performance without pruning the tree.

Model building: To build the model, ten different experiments were conducted using J48 decision tree, Naïve Bayes classifier and PART rule induction. All experimentations during model building applied after SMOTE (300%) (Table 21).

Experimentation with J48 algorithm: J48 is Weka’s implementation of the C4.5 algorithm which can work on multiple valued attributes. In addition to using the default parameter settings of the algorithm to build predictive model with J48, an attempt was made to find better classifier by varying its important parameters (Table 22) shows that J48 pruned experimentations with all attributes.

Experimentation with naïve bayes algorithm

Bayesian methods are based on the assumptions of probability. The Naïve Bayes algorithm assumes the attributes are independent.

Exp#	Experiments	Scenarios (Models)
1	J48 pruned Tree Model Generation	J48 pruned tree with all 16 attributes
2	J48 unpruned Tree Model Generation	J48 unpruned tree with all 16 attributes
3	J48 pruned Tree Model Generation	J48 pruned tree with 12 selected attributes
4	J48 unpruned Tree Model Generation	J48 unpruned tree with 12 attributes
5	Naïve Bayes Classifier	Naïve Bayes with all 16 attributes
6	Naïve Bayes Classifier	Naïve Bayes with 12 selected attributes
7	PART pruned rule induction	PART pruned with all 16 attributes
8	PART unpruned rule induction	PART unpruned with all 16 attributes
9	PART pruned rule induction	PART pruned with 12 selected attributes
10	PART unpruned rule induction	PART unpruned with 12 selected attributes

Table 21: Experiments and Scenarios.

Exp#	Model	SMOTE	Acc	WTPR	WFPR	WROC
1	J48 pruned tree with all attributes	Before	92.6%	92.6%	1.5%	0.984
2	J48 pruned tree with all attributes	100%	93.3%	93.3%	1.3%	0.985
3	J48 pruned tree with all attributes	200%	94.2%	94.2%	1.8%	0.981
4	J48 pruned tree with all attributes	300%	92.2%	92.2%	2.6%	0.973

Table 22: Experimentation with J48 Decision Tree.

Exp#	Model	SMOTE	Acc	WTPR	WFPR	WROC
1	Naïve Bayes with all attributes	Before	92.2%	92.2%	1.6%	0.986
2	Naïve Bayes with all attributes	100%	93.1%	93.1%	1.3%	0.989
3	Naïve Bayes with all attributes	200%	93.1%	93.8%	1.9%	0.984
4	Naïve Bayes with all attributes	300%	89.7%	89.7%	2.8%	0.976

Table 23: Experimentation with Naïve Bayes Classifier.

Exp#	Model	SMOTE	Acc	WTPR	WFPR	WROC
1	PART pruned rule with all attributes	Before	91.1%	91.1%	2.2%	0.978
2	PART pruned rule with all attributes	100%	92.3%	92.3%	1.3%	0.981
3	PART pruned rule with all attributes	200%	93.2%	93.2%	2%	0.98
4	PART pruned rule with all attributes	300%	92.6%	92.6%	2.5%	0.978

Table 24: Experimentation with PART rule induction.

The probability of co-occurrence of an attribute value together with a particular outcome value is computed. Then, the class of a new instance will be computed by multiplying the probabilities of values the instance has assumed under each attribute. (Table 23) depicts experimentation with Naïve Bayes with all attributes.

Experimentation with PART rule induction algorithm

PART rule induction algorithm can work on multiple valued attributes (Table 24).

Based on the experimental design, (Table 25) depicted that experimentation of three algorithms through changing their schemes.

Performance evaluation and comparison of classifiers

One of the objectives of this study was to compare and evaluate the techniques which were used in the study, such as decision tree, Naïve Bayes and PART rule induction classifiers and to select the one, which performs the best. To do this, the standard metrics of accuracy, True-Positive Rates, False-Positive Rates and ROC are applied.

The next task in testing the model to decide which one of the six models constitutes a better model/classifier of the 2011 EDHS dataset is evaluated by using ROC area analysis. ROC is the main indicator

Exp#	Model	Accuracy	WTPR	WFPR	WROC
1	J48 pruned tree with all attributes	92.24%	92.2%	2.6%	0.973
2	J48 unpruned tree with all attributes	92.73%	92.7%	2.6%	0.966
3	J48 pruned tree with selected attributes	90.58%	90.6%	2.4%	0.975
4	J48 unpruned tree with selected attributes	91.68%	91.7%	2.9%	0.973
5	Naïve Bayes with all attributes	89.68%	89.7%	2.9%	0.976
6	Naïve Bayes with selected attributes	89.94%	89.8%	2.8%	0.976
7	PART pruned rule with all attributes	92.62%	92.6%	2.5%	0.978
8	PART unpruned rule with all attributes	92.82%	92.8%	2.27%	0.96
9	PART pruned rule with all attributes	91.72%	91.7%	2.8%	0.976
10	PART unpruned rule with all attributes	90.84%	90.8%	3.3%	0.971

Table 25: Experimentation with three selected algorithms.

during algorithm performance selection. In this study, imbalanced data balanced using SMOTE analysis. For this reason, ROC area is the best indicator for SMOTEd data rather than accuracy. With regards to ROC area, a model with perfect accuracy will have an area of 1.0 i.e. the larger the area, the better performance of the model or the larger values of the test result variable indicates the stronger evidence for a positive actual state. ROC analysis in which the curve the more to the upper left would indicate a better classifier. Here in this case, the WROC area performance of the algorithms show that PART rule induction, Naïve Bayes with all attributes and J48 pruned decision tree algorithm with all attributes scored the highest area of 97.8% and 97.6% and 97.3% respectively. The lowest WROC keeps account in PART unpruned rule induction with all attributes which is 96%.

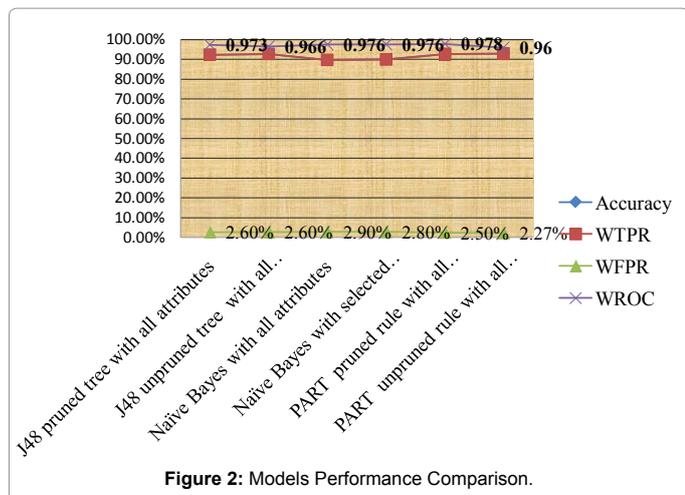
Selected model performance and evaluations

During PART pruned rule induction model generation, the effect of the attributes on the model performance was investigated. The full training set containing a total of 14,501 instances were used in all attributes. In addition to the above performance metrics (accuracy, WTPR, WFPR and WROC area) used, relatively PART pruned rule induction with all attributes is more understandable and less complex to human than others model generated. Therefore, the performance of PART pruned rule induction classifier with all attributes gives valuable information in predicting nutritional status as compared to other models (Figure 2).

Results

In this study, six experiments have been conducted using three data mining classification algorithms i.e. J48 algorithm, Naïve Bayes and PART rule induction classifier in order to build a model that predicts nutritional status of under-five years children in Ethiopia. These three algorithms have literature support. Models developed using J48 decision tree, Naïve Bayes classifier and PART rule induction algorithms have good accuracy and ROC results during test experiments. Based on these the experiments were designed for four purposes; to investigate the effect of tree pruning methods when building a decision tree model, to observe how attribute selection affects the classification accuracy, to compare J48 decision tree, Naïve Bayes and PART rule induction classifier and to extract significant rules. During test with selected 12 attributes accuracy in J48 and PART rule classifiers showed slight decrease as compared to all 16 attributes results. Due to this Naïve Bayes experiments done using with all and selected attributes.

With regards to effect of pruning, it is obvious that model with grown size of tree make the model difficult to understand and



interpret by human as well as generating the rule become challenging. Experiments were done to reduce the complexity of the tree so as to make model more compact and understandable. Therefore, the models are experimented through pruning the tree on the training schemes.

In this study, the model created using PART pruned rule induction classifier registers good performance (i.e 97.8% WROC area) and hence selected for further analysis/rule tracing.

Rule extraction

To make decision tree model and PART decision list model more human-readable each path from root to leaf can be transformed into an IF-THEN rule. If the condition is satisfied, the conclusion follows. PART rule induction algorithm is the best known method for deriving rules from classification lists. Both PART rule induction and J48 decision tree classifiers follow decision lists, IF-THEN rule such as PART rule decision list:

CHILD ANEM=Not anemic AND, MOTH OCCUP=Working AND, MOTH EDUC=No education AND, TOTAL=>6 AND, SEX=Female AND, MOTH BMI=<18.5: Stunted (17.0/1.0)

The numbers in (parentheses) indicates the number of examples in the leaves. The number of misclassified examples would also be given, in this case 1 (6%) after a slash (/) and hence it is possible to compute the success fraction (ratio) to estimate the level of confidence or likelihood of predictability of the class that tells how much the rule is strong.

PART pruned rule induction model with all attributes produced 346 different rules. However, the researcher selected best interesting rules that cover most of the data points in the study. The other things for selection of rule are new finding and the ability of classifying large instances correctly.

Some specific rules extracted for “Stunted” class:

Rule#1: If HAZ=<-2SD AND WAZ=-2SD to 2SD: Then the class Stunted (3233.0/8.0)

This rule revealed 3233(99.7%) correctly out of 3241 instances.

Rule#2: If HAZ=<-2SD AND WHZ=-2SD-2SD AND RESIDENCE=Rural AND

CHILD ANEM (child anemia level)=Mild AND MOTH EDU=Primary AND TOTAL

(Total number of children)=5-6 AND SIZE (child size at birth)=Normal: Then the

Class Stunted (12.0/1.0)

Rule#3: If HAZ=<-2SD AND CHILANEM= Mild AND RESIDENCE=Rural AND

MOTH BMI (mother’s body mass index)=<18.5 AND CHILAGE=12-23 AND

SIZE =Small: Then the class Stunted (12.0/3.0)

Rule#4: If HAZ=<-2SD AND CHILANEM=Mild AND RESIDENCE=Rural AND

WEALTHINDEX=Middle AND MOTHEduc=No education AND MOTHAGE= 25-29: Then the class Stunted (12.0/3.0)

Rule#6: If HAZ=<-2SD AND CHILANEM=Moderate AND RESIDENCE=Rural AND

MOTHBMI=<18.5 AND EVER HAD VAC=Yes AND WEALTH INDEX=Poorest

AND REGION=Afar: Then the class Stunted (17.0/1.0)

Rule#7: If HAZ=<-2SD AND CHILD ANEM=Moderate AND MOTH EDUC=No education

AND MOTH BMI=18.5-24.9AND MOTH AGE=25-29 AND TOTAL=3-4 AND

SIZE=Normal AND WEALTH INDEX=Poorest: Then the class Stunted (7.0/2.0)

The IF-THEN shows rules generated by PART pruned rule induction model to classify nutritional status as stunted (long term or chronic malnutrition). In the case of rule 1, capability of test classification, 3233 (99.7%) correctly and 8 (0.03%) instances classified incorrectly with HAZ attributes (where HAZ<-2SD). HAZ alone matter the fate of the child malnourished as stunted keeping other variables constant. In addition to rule1, rule 2 revealed that total number of children in a household, mother’s education, mother’s body mass index and residence become indicators for child’s nutritional status with correct classification of 92.3% instances. Domain area experts and literatures also support these rules [5].

Child age category becomes predictors for stunting with combination of height, i.e., HAZ. In the PART pruned decision list, HAZ would become the first indicator for stunting. Children in rural areas are likely to be stunted and regional variation in the prevalence of stunting in children is substantial. The mother’s nutritional status, as measured by her body mass index (BMI), also has a relationship with her child’s level of stunting in the case of rules 3 and 6. Relationship is also observed between the household wealth index and the stunting levels of children in rules 6 and 7.

Some specific rules extracted for “Underweight” class:

Rule#1: If WAZ=<-2SD AND WHZ=-2SD-2SD AND REGION=Benishangul-Gumuz AND SEX =Female AND MOTH BMI=<18.5: THEN the class Underweight (71.0/1.0)

Rule#2: If WAZ=<-2SD AND WHZ=-2SD-2SD AND REGION=SNRP AND RESIDENCE=Rural AND SIZE=Small AND CHILD AGE=12-23: THEN the class Underweight (18.0)

Rule#3: If WAZ=<-2SD AND WHZ=-2SD-2SD AND REGION=Benishangul-Gumuz AND CHILD ANEM=Moderate AND

MOTH EDUC=No education: THEN the class Underweight (49.0/3.0)

Rule#4: If WAZ=<-2SD AND WHZ=-2SD-2SD AND CHILD ANE=Mild AND RESIDENCE=Rural AND WEALTH INDEX=Poorest AND REGION= SNNP AND MOTH EDUC= No education: THEN the class Underweight (30.0)

Rule#5: If WAZ=<-2SD AND CHILD ANEM=Not anemic AND CHILD AGE=36-47 AND MOTH AGE=25-29 AND MOTH OCCUP=Not working AND SIZE=Small:

THEN the class Underweight (25.0/1.0)

The above rules extracted by PART pruned rule induction model applied to predict nutrition status as underweight. The capability of correctly identifying instances is good. In rule 1, when WAZ<-2SD and with others related attributes extracted 71 (98.2%) correctly and 1 (1.8%) incorrectly. Rules reveal that underweight children are experienced in the age groups 12-23 and 36-47 months. This may be explained by the fact that foods for weaning are typically introduced to children in the older age group, thus increasing their exposure to infections and susceptibility to illness. This tendency, coupled with inappropriate or inadequate feeding practices, may contribute to faltering nutritional status among children in these age groups. Being small size at birth has likely to be underweight later in life. According to rule 1, children born to mothers who are thin (BMI less than 18.5) are more likely to be underweight. The proportion of underweight children is higher for those born to uneducated mothers. Rule 4 shows that underweight children decrease as the wealth quintile of the mother increases [14].

Some specific rules extracted for “Wasted” class:

Rule#1: If WHZ=<-2SD AND WAZ =-2SD-2SD: Then the class Wasted (635.0/3.0)

Rule#2: If WHZ=<-2SD AND CHILDANEM=Mild AND MOTHEDUC=No education: Then the class Wasted (205.0)

Rule#3: If WHZ=<-2SD AND TOTAL=>6 AND MOTHBMI=18.5-24.9: Then the class Wasted (120.0/2.0)

Rule#4: If WHZ=<-2SD AND HAZ=-2SD- 2SD AND SEX=Male: Then the class Wasted (45.0)

Rule#5: If WHZ=<-2SD AND REGION=Amhara AND SEX=Male AND MOTHBMI=18.5-24.9 AND CHILDANEM=Moderate AND MOTHOCUP=Not working: Then the class Wasted (7.0)

Rule#6: If WHZ=<-2SD AND REGION=Dire Dawa: Then the class Wasted (101.0/2.0)

Rule#7: If WHZ=<-2SD AND REGION=Afar AND WEALTHINDEX=Poorest AND TOTAL=5-6: Then the class Wasted (85.0/1.0)

The above decision list rules predicted as wasted (acute malnutrition) by the PART pruned rule induction classifier. Rule 1 revealed that the probability of being wasted is most predicted (99.5%) when WHZ is less than minus two standard deviations keeping all variables constant (WHZ<-2SD). Rules 4 and 3 showed that wasting is slightly more likely in male than female children. This showed that currently a child would be diarrheic or/and not breastfeed accordingly. The rule revealed that wealth index of the household is between poorer and rich whose mother has primary education a child become wasted due to not working or lack of health service utilization awareness of mother's.

Some specific rules extracted for “Normal” class:

Rule#1: If HAZ=-2SD- 2SD AND WAZ=-2SD-2SD AND WHZ=-2SD-2SD AND

RESIDENCE=Rural: Then the class Normal (3472.0/1.0)

Rule#2: If HAZ=-2SD- 2SD AND WAZ=-2SD-2SD AND WHZ=-2SD-2SD AND RESIDENCE=Urban AND TOTAL=1-2 AND REGION=Addis Ababa AND

SIZE=Normal: Then the class Normal (72.0)

Rule#3: If HAZ=-2SD- 2SD AND WAZ=-2SD-2SD AND WHZ=-2SD-2SD AND

RESIDENCE=Urban AND TOTAL=1-2 AND REGION=Addis Ababa AND

SIZE=Small AND MOTHOCUP=Working AND SEX=Male: Then the class

Normal (8.0)

Rule#4: If HAZ=-2SD- 2SD AND WAZ=-2SD-2SD AND WHZ=-2SD-2SD AND

RESIDENCE=Urban AND TOTAL=1-2 AND REGION=Somali: Then the class

Normal (19.0)

As shown in rule#1, IF-THEN decision list confirmed the three anthropometric measurements, i.e. HAZ, WAZ and WHZ are between minus two and positive two standard deviation (-2SD to 2SD) of those children who have good or normal nutritional status. The first rule gave a correct result for 3472 of the 3473 instances, thus its success fraction is 3472/3473. Rule 1 (when HAZ, WAZ, WHZ=-2SD to 2SD) indicates that the likelihood of a child being at a good nutritional status, keeping all predictors of the class normal is 99.97%.

It is expected that mother's education is likely to improve nutritional status of children through better use of health facilities and better child care practices. Moreover, mother's occupation for her family income source, mother's nutritional status, residence and total number of children need interventions because of in almost all rules they have important prediction values

Error rate of the selected model

In classification or prediction tasks, the accuracy of the resulting model is measured either in terms of the percentage of instances correctly classified or in terms of “error rate” i.e. the percentage of records classified incorrectly. Classification error rate on pre classified test set is commonly used as an estimate of the expected error rate when classifying new records [15]. Errors during each test are averaged to give the average error rate of the model. The classification error rate for the selected model is 7.38%, which means the model has incorrectly classified about 7.38% instances out of their actual classes each time when the model is tested on the test set.

The percentage of incorrectly classified instances indicates the chance with which the developed model misclassifies a new victim out of the actual class. Several reasons may be attributed for increased error rate from the models. First, algorithms differ in their capability as observed from comparisons of performance measures. Second, attributes may not be included in the collection and study might have influenced it. All the models of the predictive performance in identifying True Positive cases of model are higher than identifying

True Negative cases. Consequently, the model tends to misclassify instances to some other classes.

Conclusion and Recommendation

Application of data mining technology has increasingly become very popular and proved to be relevant for many sectors such as healthcare sector, has been applied for patient survival analysis, prediction of diagnosis, for outcomes measurement, to improve patient care and decision-making etc. However, the potentials of data mining have not yet been used in predicting nutritional status of under-five children in Ethiopia. In this study, the objective was to design a predictive model for nutritional status of under-five children using data mining techniques using 2011 EDHS dataset. The model would be used in the future so as to help policy makers and health care providers in the country to identify children who are at risk. Furthermore, such a predictive model might be applied in assisting under-five malnutrition prevention and control activities in the country.

The hybrid, iterative methodology, was employed in this study which consists of six basic steps such as problem domain understanding, data understanding, data preparation, data mining, and evaluation of the discovered knowledge and use of the discovered knowledge.

In order to generate interesting rule from the huge data collected in the 2011 EDHS, a total of 9607 instances and 17 attributes were applied. Knowledge discovery in dataset was employed after SMOTE technique was applied which is an automatic operation where minority classes are over sampled to make the target attribute balance.

In this particular research, independent variables/attributes were: mother's age, region, residence, mother's occupation, mother's education, mother's BMI, wealth index, total number of children, age of child, sex of child, ever had vaccination, size of child at birth, child anemia level, HAZ, WAZ, WHZ and dependent variable/attributes was under nutrition status.

The findings clearly suggested that most of the above attributes had strong relation with nutritional status of under-five children in the demographic and health survey data. All the selected attributes were used in the analysis using J48 decision tree, Naïve Bayes and PART rule induction algorithms.

Several models were built during experimentation that could predict the risk of under-five malnutrition. Among these models, PART pruned rule induction model with all attributes showed an interesting predictive accuracy result of 92.6% and ROC area of 97.8% (Figure 3).

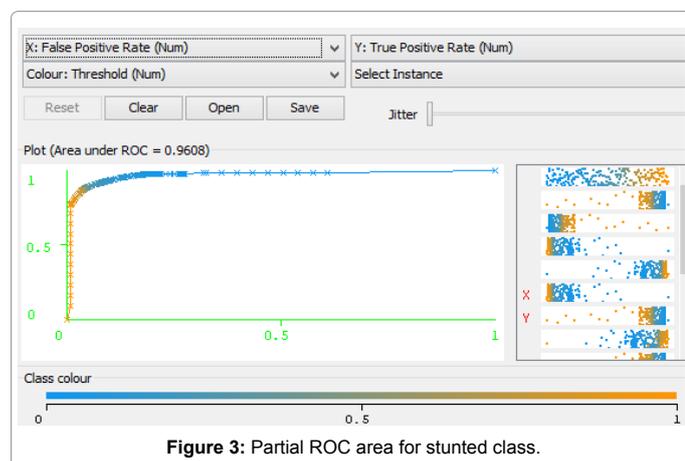


Figure 3: Partial ROC area for stunted class.

In summary, region, residence, mother's education, wealth index of the household, child age are major contributor of under nutrition and predictive model is developed using PART pruned rule induction with all attributes to predict nutritional status of under-five children in Ethiopia.

Recommendations

In this study, an attempt was made to explore the 2011 EDHS dataset and to provide an initial insight into the potential applicability of data mining techniques in predicting nutritional status of under-five children based on demographic, health and socioeconomic characteristics. Reducing child malnutrition and improving child health status, through appropriate interventions, requires better understanding of the main demographic, health and socioeconomic determinants. Thus, based on the result of the research learned by PART pruned rule induction algorithm; the following recommendations were made by the researchers.

➤ Models presented in this study showed that under nutrition can be reduced substantially by intervening in certain socio-economic and demographic factors so that probability of under-five malnutrition can be minimized. Thus models should be used for formulating child nutrition programs and child health policies.

➤ It has been observed that developing many other classifiers for prediction with the short period of time given to this research was unlikely. Therefore, to enhance the performance of the present model further research should be conducted on nutritional status of under-five area incrementally using many more mining techniques to improve the predictive model accuracy.

➤ This research developed predictive model for under five children nutritional status prediction. However, there is a need for knowledge based system for predicting under five children nutritional status. Therefore, organizations working on under-five children nutrition should also work on the determinant identified findings.

➤ This study considered demographic and health survey dataset to predict nutritional status of under-five children. So that future studies might need to discover knowledge and patterns in other domain areas such as clinical datasets.

➤ All the predictors managed through using different strategy such as giving health education on nutrition, childcare, family planning at community based through rural health extension workers at rural area will improve nutritional status of under-five children. Therefore, health professionals who are working on nutrition should strengthen appropriate health intervention.

Competing Interests

The authors declared that they have no competing interests.

Authors' Contributions

Zenebe Markos wrote the proposal, participated in data collection, analyzed the data and drafted the paper. Dr. Martha Yifiru and Dr. Jemal Haidar approved the proposal with some revisions, participated in data collection and analysis, commented on the analysis and improved the first draft. All the three authors and Feleke Doyore revised subsequent drafts of the paper. Feleke Doyore prepared this manuscript for publication.

Acknowledgment

Our earnest gratitude goes to Health and Medical sciences college, Addis

Ababa University for proper review and approval of this paper. We would also like to extend our gratitude to data collectors for their patience to bring this meaningful information. Our special thanks also extended to Addis Ababa University and Hossana College of health sciences for financial support for this study.

References

1. UNESCO Report (2012) The Review of Health and Nutrition indicators in early childhood.
2. Ethiopia Federal Ministry of Health (2007) Protocol for the management of severe acutemalnutrition, Ethiopia, Addis Ababa.
3. (2012) Tackling child malnutrition, Save the Children.
4. Aweke K, Habtamu F, Akalu G (2012) Nutritional status of children in food insecure households in two districts of north showa zone, Ethiopia, 12: 2.
5. (2012) Central Statistical Agency, Addis Ababa: Ethiopia Demographic and Health Survey 2011, Ethiopia ICF International Calverton, Maryland, USA 5.
6. Mehamed Kantardzic JB (2003) Data Mining-Concepts, Models, Methods, and Algorithms. USA: John Wiley & Sons Publication Inc 6.
7. (2005) Two Crows Corporation, Introduction to Data Mining and Knowledge Discovery 3rd ed. Two Crows Corporation. 500 Falls Road, Potomac, USA.
8. Cios Krzysztof J, Pedrycz Witold, Swiniarski Roman W, Kurgan Lukasz A (2007) Data Mining: A Knowledge Discovery Approach. New York, USA: Springer Science Business Media LLC, 8.
9. Colin Shearer (2000) The CRISP-DM Model: The New Blueprint for Data Mining. *Journal of Data Ware Housing* 5: 4.
10. (2011) Ethiopian Demographic and Health Survey, Central Statistics Agency Addis Ababa, Ethiopia, ICF International, Calverton, Maryland, USA.
11. Han Jiawei, Kamber Micheline (2001) Data Mining: Concepts and Techniques. New York. USA: Morgan Kaufmann Publishers 11.
12. David L Olson, Dursun D (2008) Advanced Data Mining Techniques. Springer-Verlag Berlin Heidelberg 12.
13. Weiss Sholom M, Zhang Tong (2003) Performance Analysis and Evaluation. In:Ye Nong, Editor. The Hand Book of Data Mining. New Jersey. USA: Lawrence Erlbaum Associates Inc 13.
14. Abdu Kedir Seid (2012) Health and Nutritional Status of Children in Ethiopia: Do Maternal Characteristics Matter? *J Biosoc Sci* 45: 187-204.
15. Bath Peter A (2004) Data Mining in Health and Medical Information: Annual review of Information Science and Technology. Blaise Cronin, 38. USA: Information Today Inc.